Plant identification: Man vs. Machine

Pierre Bonnet, Alexis Joly, Hervé Goëau, Julien Champ, Christel Vignau, Jean-François Molino, Daniel Barthélémy, Nozha Boujemaa

Noname manuscript No.
(will be inserted by the editor)

Plant identification: Man vs. Machine

LifeCLEF 2014 plant identification challenge

Pierre Bonnet · Alexis Joly · Hervé Goëau · Julien Champ · Christel Vignau · Jean-François Molino · Daniel Barthélémy · Nozha Boujemaa

Received: date / Accepted: date

Abstract This paper reports a large-scale experiment aimed at evaluating how state-of-art computer vision systems perform in identifying plants compared to human expertise. A subset of the evaluation dataset used within LifeCLEF 2014 plant identification challenge was therefore shared with volunteers of diverse expertise, ranging from the leading experts of the targeted flora to inexperienced test subjects. In total, 16 human runs were collected and evaluated comparatively to the 27 machine-based runs of LifeCLEF challenge. One of the main outcomes of the experiment is that machines are still far from outperforming the best expert botanists at the image-based plant identification competition. On the other side, the best machine runs are competing with experienced botanists and clearly outperform beginners and inexperienced test subjects. This shows that the performances of automated plant identification

P. Bonnet
CIRAD, UMR AMAP, France, E-mail: pierre.bonnet@cirad.fr

A. Joly, J. Champ
LIRMM, France, E-mail: alexis.joly@inria.fr

H. Goëau, A. Joly
Inria ZENITH team, France, E-mail: herve.goeau@inria.fr

J. Champ
INRA UMR AGAP, France, E-mail: julien.champ@lirmm.fr

C. Vignau
Tela Botanica, France, E-mail: christel@tela-botanica.org

J.-F. Molino
IRD, France, E-mail: jean-francois.molino@ird.fr

D. Barthélémy
CIRAD, BIOS Direction and INRA, UMR AMAP, F-34398, France, E-mail: daniel.barthelemy@cirad.fr

N. Boujemaa
INRIA, Direction of Saclay Center, France, E-mail: nozha.boujemaa@inria.fr

systems are very promising and may open the door to a new generation of ecological surveillance systems.

Keywords visual plant identification · human evaluation · digital data · image analysis

1 Introduction

Building accurate knowledge of the identity, geographic distribution and uses of plants is essential for a sustainable development of agriculture as well as for biodiversity conservation. Unfortunately, such basic information is often only partially available for professional stakeholders, teachers, scientists and citizens, and often incomplete for ecosystems that possess the highest plant diversity. Speeding up the collection and integration of raw botanical observation data is therefore a crucial challenge. One of the major problems towards this objective, sometimes referred as the taxonomic gap, is that identifying plant species is usually impossible for the general public, and often a very difficult task for professionals, such as farmers or foresters. On the other side, the number of experienced botanists is decreasing significantly so that collecting massive sets of plant observations with a valid taxonomic name becomes more and more challenging.
In this context, content-based image retrieval and computer vision approaches are considered as one of the most promising solutions to help bridging the gap, as discussed in [1,2,3,4,5]. We therefore see an increasing interest in this transdisciplinary challenge in the multimedia community (e.g. in [6,7,8,9,10, 11]). Beyond the raw identification performances achievable by state-of-the-art computer vision algorithms, recent visual search paradigms actually offer much more efficient and interactive ways of browsing large flora than standard field guides or online web catalogs [12]. Smartphone applications relying on such image-based identification services are particularly promising for setting-up massive ecological monitoring systems, involving thousands of contributors at a very low cost.
A first step in this way has been achieved by the US consortium behind LeafSnap[1], an i-phone application allowing the identification of 184 common american plant species based on pictures of cut leaves on an uniform background (see [13] for more details). Then, the French consortium supporting Pl@ntNet [5] went one step further by building an interactive image-based plant identification application that is continuously enriched by the members of a social network specialized in botany. Inspired by the principles of citizen sciences and participatory sensing, this project quickly met a large public with more than 200K downloads of the mobile applications [14,15]. A related initiative is the plant identification evaluation task organized since 2011 in the context of the international evaluation forum CLEF[2] ; the task is based on the data

[1] `http://leafsnap.com`
[2] `http://www.clef-initiative.eu`

collected within Pl@ntNet. In 2011, 2012 and 2013 respectively 8, 11 and 12 international research groups participated in this large collaborative evaluation by benchmarking their image-based plant identification systems (see [16, 17,18] for more details). The data used during these 3 first years can be accessed online[3]. Contrary to previous evaluations reported in the literature, the key objective was to build a realistic task very close to real-world conditions (different users, areas, periods of the year, important species number, etc.). This paper is directly in the continuity of this initiative as it builds on the analysis of the 2014th edition of the task organized with the newly created lab of CLEF LifeCLEF[4]. 74 research groups worldwide registered and 10 of them crossed the finish line by submitting runs (from 1 to 4 depending on the teams). Details of the participants and the methods used in the runs are synthesised in the overview working note of the task [19]. The main contribution of this new paper is to complete the system-oriented evaluation of LifeCLEF by a user trial aimed at comparing the performances of state-of-the-art automatic recognition systems with the human expertise (see section 3). A random selection of observations of the test set was realised and submitted to a panel of volunteers in the aim to allow strict comparison between computer vision systems and human expertise. Additionally (and complementary), we also extended the raw official results of the benchmark by a deeper analysis of the runs with regard to the image quality of the test observations (see section 2).

2 Analysis of LifeCLEF plant identification task results

2.1 Dataset

The data used for the plant task of LifeCLEF (and all previous plant tasks of ImageCLEF) is collected through TelaBotanica[5], the main French-speaking botany network linking 20K registered members living worldwide in more than 70 countries. About 41% of the members declare being novice in botany, while 30% declare having a good practice and 7% to be experts. 53% of the members have a professional activity related to botany. The main activities of the network are organized around tens of collaborative projects coordinated by experts and dedicated to distinct botany issues (plant usages, distribution, identification, specific flora, etc.). One of them was specifically launched in 2010 to collect training images to be used for automatic identification purposes. The botanical target of the project has been gradually increased from leaf scans of French trees at the beginning to the whole French flora and multiple views of plants nowadays. The expected images are defined by an acquisition protocol that is disseminated to the members of the project through illustrated booklets containing full-text descriptions as well as positive and negative examples of the different view types (full plant, flower, leaf, fruit, bark, leaf scan).

[3] http://publish.plantnet-project.org/project/plantclef

[4] http://www.lifeclef.org

[5] http://www.tela-botanica.org

Once raw image data of plants have been collected, they are integrated through a collaborative tagging tool allowing to enter the taxon's name and view type of each observation, as well as complementary metadata (area, author, quality rating, project identifier, etc.). It is possible to upload unidentified images of plants so that other members of the social network can determine the right species. Besides, contradictory determinations can be added to the observations posted and identified by other users. Each image can therefore be associated to several determinations made by distinct members and the best consensus is computed by voting. People who made contradictory determinations can discuss through a forum associated to each ambiguous image and the determinations can be revised anytime. Finally, the quality of each image with respect to a given protocol can also be entered by several users through a rating box. For the LifeCLEF evaluation campaign, observations were considered as valid only if (i) at least two persons annotated the content (ii) there is no unresolved conflict on the determination. We finally kept only the 500 species with the more observations (the minimum being 5 observations).
The resulting PlantCLEF 2014 dataset is composed of 60,962 pictures belonging to 19,504 observations of 500 species of trees, herbs and ferns living in Western Europe (present in France, and neighboring countries). This data was collected by 1,608 distinct contributors. Each picture belongs to one and only one of the 7 types of views reported in the meta-data (entire plant, fruit, leaf, flower, stem, branch, leaf scan) and is associated with a single plant observation identifier allowing to link it with the other pictures of the same individual plant (observed the same day by the same person). It is noticeable that most image-based identification methods and evaluation data proposed in the past were so far based on leaf images (e.g. in [13, 20, 21] or in the more recent methods evaluated in [17]). Only few of them were focused on images of flower as in [6] or [33]. Leaves are far from being the only discriminant visual key between species but, due to their shape and size, they have the advantage to be easily observed, captured and described. More diverse parts of the plants however have to be considered for accurate identification. As an example, the 6 species depicted in Figure 1 share the same French common name of *"laurier"* even though they belong to different taxonomic groups.

The main reason is that these shrubs, often used in hedges, share leaves with more or less the same-sized elliptic shape. Identifying a *laurel* can be very difficult for a novice by just observing leaves, while it is indisputably easier with flowers. Beyond identification performances, the use of leaves alone has also some practical and botanical limitations. Leaves are not visible all over the year for a large fraction of plant species. Deciduous species, distributed from temperate to tropical regions, can't be identified by the use of their leaves over different periods of the year. Leaves can be absent (i.e. leafless species), too young or too degraded (by pathogen or insect attacks), to be exploited efficiently. Moreover, leaves of many species are intrinsically not informative enough or very difficult to capture (needles of pines, thin leaves of grasses, huge leaves of palms, ...).
Another originality of the PlantCLEF dataset is that its social nature makes

Fig. 1 6 plant species sharing the same common name for *laurel* in French, belonging to distinct species.

it closer to the conditions of a real-world identification scenario: (i) images of the same species are coming from distinct plants living in distinct areas (ii) pictures are taken by different users that might not have used the same protocol to acquire the images (iii) pictures are taken at different periods in the year. Each image of the dataset is associated with contextual meta-data (author, date, locality name, plant id) and social data (user ratings on image quality, collaboratively validated taxon names, vernacular names) provided in a structured xml file. The gps geo-localization and the device settings are available in many images.
Table 2 gives some examples of pictures with decreasing averaged user ratings for the different types of views. Note that the users of the specialized social network creating these ratings (Tela Botanica) are explicitly asked to rate the images according to their plant identification ability and their accordance to the pre-defined acquisition protocol for each view type. This is not an aesthetic or general interest judgement as in most social image sharing sites.

2.2 Task Description

The task was evaluated as a plant species retrieval task based on multi-image plant observations queries. The goal is to retrieve the correct plant species among the top results of a ranked list of species returned by the evaluated system. Contrary to previous plant identification benchmarks, queries are not defined as single images but as *plant observations*, meaning a set of one to several images depicting the same individual plant, observed by the same person, the same day. As illustrated in Figure 3, each image is associated with one view (entire plant, branch, leaf, fruit, flower, stem or leaf scan) and with contextual meta-data (data, location, author). Semi-supervised and interactive approaches were allowed but as a variant of the task and therefore evaluated independently from the fully automatic methods. None of the participants, however, used such approaches in the 2014 campaign.

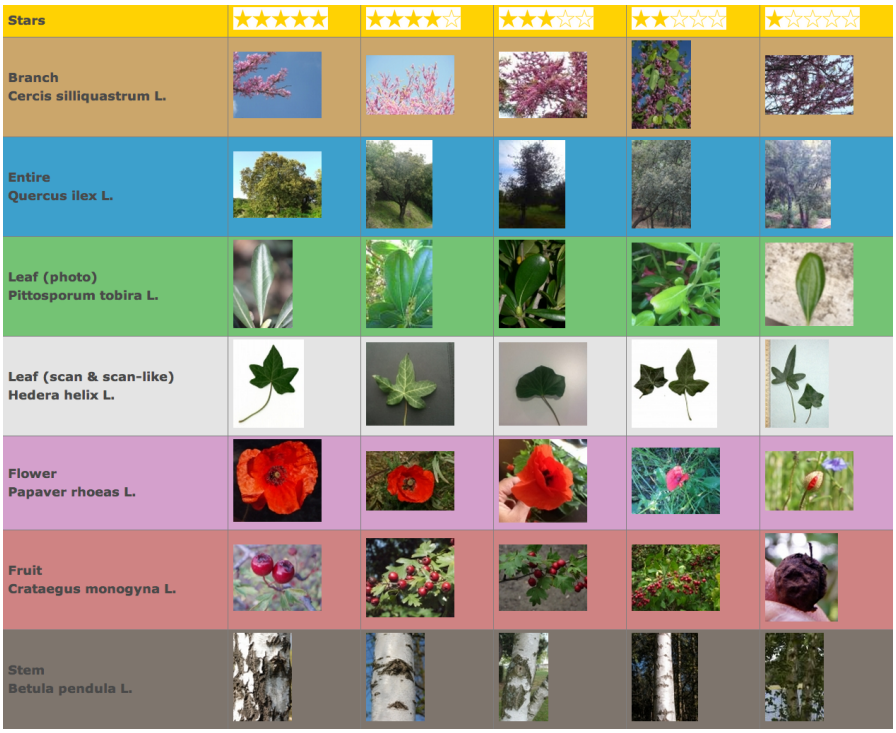

Fig. 2 Examples of PlantCLEF pictures with decreasing averaged users ratings for the different types of views

In practice, the whole PlantCLEF dataset was split in two parts, one for training (and/or indexing) and one for testing. The training set was delivered to the participants in January 2014 and the test set two months later so that participants had some time to become familiar with the data and train their systems. After the delivery of the test set, participants had two additional months to run their system on the undetermined plant observations and finally send their results files. Participants were allowed to submit up to 4 distinct runs. More concretely, the test set was built by randomly choosing 1/3 of the observations of each species whereas the remaining observations were kept in the reference training set.

This resulted in a dataset presented in Table 2.2. The xml files containing the meta-data of the *query* images were purged so as to erase the taxon name (the ground truth) and the image quality ratings (that would not be available at query stage in a real-world mobile application). Meta-data of the observations in the training set were kept unaltered.

In practice, each candidate system is then evaluated through the submission of a *run*, i.e. a file containing a set of ranked lists of species (each list corresponding to one query observation and being sorted according to the

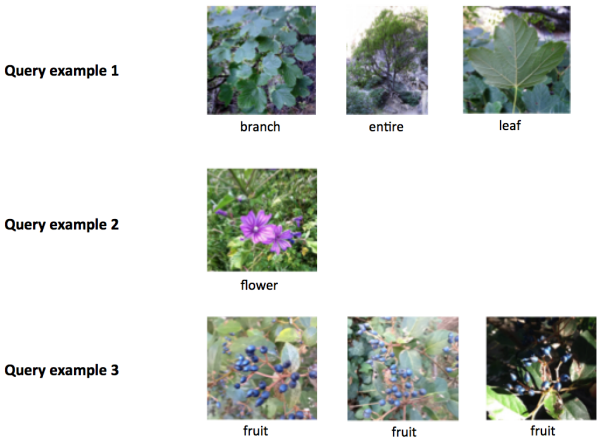

Fig. 3 Query examples – some of them are composed of a single picture whereas some others contain several pictures of one or several types of view

Dataset	#images	Branch	Entire	Flower	Fruit	Leaf	Stem	Scan & scan-like
Train	47815	1987	6356	13164	3753	7754	3466	11335
Test	13146	731	2983	4559	1184	2058	935	696

Table 1 Statistics on the number of images by datasets, and by image types

confidence score of the system in the suggested species). The metric used to evaluate the submitted runs is an extension of the mean reciprocal rank [34] classically used in information retrieval. The difference is that it is based on a two-stage averaging rather than a flat averaging such as:

$$S = \frac{1}{U} \sum_{u=1}^{U} \frac{1}{P_u} \sum_{p=1}^{P_u} \frac{1}{r_{u,p}} \tag{1}$$

where U is the number of users (within the test set), P_u the number of individual plants observed by the u-th user (within the test set), $r_{u,p}$ is the rank of the correct species within the ranked list of species returned by the evaluated system (for the p-th observation of the u-th user). Note that if the correct species does not appear in the returned list, its rank $r_{u,p}$ is considered as infinite. Overall, the proposed metric allows compensating the long-tail distribution effects of our social data. As any social network, few people actually produce huge quantities of data whereas a vast majority of users (the long tail) produce much less data. If, for instance, only one person did collect an important percentage of the images, the classical mean reciprocal rank over a random set of queries will be strongly influenced by the images of that user to the detriment of the users who only contributed with few pictures. This is a problem for several reasons: (i) the persons who produce the more data are

usually the most expert ones but not the most representative of the potential users of the automatic identification tools. (ii) The large number of the images they produce makes the classification of their observations easier because they tend to follow the same protocol for all their observations (same device, same position of the plant in the images, etc.) (iii) The images they produce are also usually of better quality so that their classification is even easier.

2.3 Quality-wise results

A total of 10 participating groups submitted 27 runs that are summarised in the overview working note of the task [19] and further developed in the individual working notes of the participants who submitted one (BME TMIT [22], FINKI [28], I3S [31], IBM AU [26], IV-Processing [30], MIRACL [23], PlantNet [24], QUT [25], Sabanki-Okan [27], SZTE [29]). Official results of the evaluation (based on the score S detailed above) are synthesized in the overview paper of LifeCLEF lab [32] and further developed in the working note of the plant task [19]. For the purpose of this paper, we refined the analysis of the collected runs in order to assess the impact of the quality of the test pictures on the identification performances of the evaluated systems. We therefore split the test observations in five categories according to the rounded average number of stars of the images composing it. We remind that the number of stars attributed to each image is itself an average number of the quality ratings provided by the members of Tela Botanica social network and supposed to reflect the interest of its visual content in terms of identification. We then re-computed a per-category score S for each run and each quality category. These new results are reported on Figure 4 and **??** for the 15 best runs of the challenge. Note that the initial ranking of the runs has been preserved in accordance to the overall official score of each run. This allows analysing if the per-category rankings of the methods differ from that global one (and/or between each other). Besides, the graph allows checking whether the overall performance of a given run is stable or strongly affected by the query image quality.

A first overall conclusion is that the performances of all runs degrade with the average quality of the observations. The scores achieved on high-quality pictures (with 4 or 5 stars) are two or three times the ones achieved on low-quality pictures (with 1 or 2 stars). Whatever the underlying methods (e.g. Fisher vectors and support vector machines, deep convolutional neural networks, local features and instance-based classifier, etc.), the impact of the quality is high in all runs. This confirms that the human ratings accorded to the visual content of the pictures are very well correlated with the ability of the algorithms to recognize the captured species. Note that this was not obvious for several reasons including the fact that some categories are more populated than others in the dataset or that some teams could have implemented specific quality-aware training strategies. The interesting point of the sensitivity to image quality is that the performances on high-quality pictures

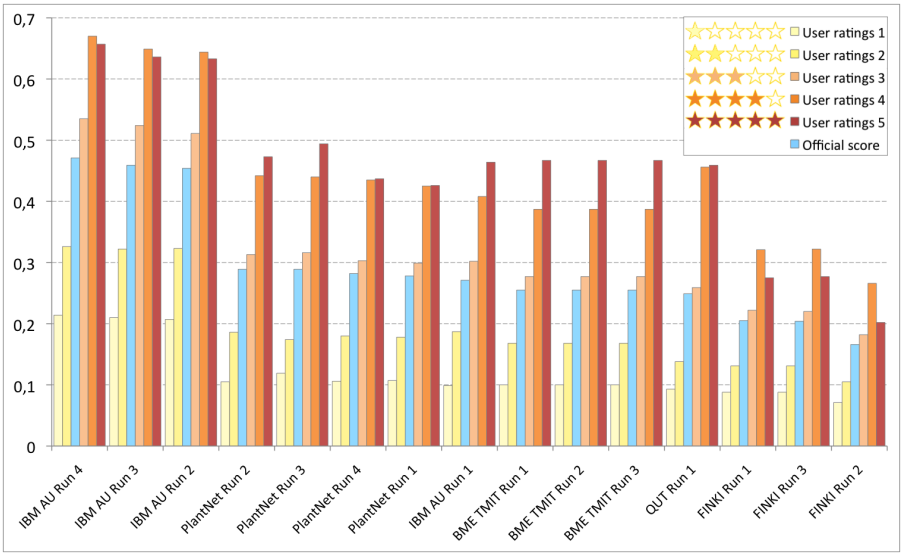

Fig. 4 Identification performances of the 15 best runs detailed by quality categories

are also consistently better than the average performances. That means that it is highly beneficial to teach potential users to take good pictures if they would like to maximize their chance to identify a given plant observation. Note that this is not a matter of taking professional pictures with expensive devices but rather a matter of respecting a simple protocol and taking a few pictures until one is good enough. All pictures in the 1 and 2 stars categories do for instance not follow the basic instructions provided within the citizen science projects launched by Pl@ntNet [5] (i.e. having one single part of the plant roughly centered in the image). Pictures in the 1 star category are usually so bad that any user understands that he cannot expect good results from it.

Now, looking into the details, we can see that all methods are not equally affected by the image quality. We can first observe that the overall ranking of the runs is modified across the categories. For instance the IBM AU Run 1, the 3 runs of the BME TMIT team and the QUT Run 1 obtained better scores than the PlantNet 2, PlantNet 4 and PlantNet 1 runs for the highest quality category 5, while the overall ranking of the runs is more or less preserved for the lowest quality categories 1, 2 & 3. The QUT Run 1 reach the 4th best score for the quality category 4. This means that the methods used by BME TMIT, QUT and IBM AU in these runs are more sensitive to the quality of the visual content of the pictures, especially for the QUT Run 1 were the differences between the 5 categories are amplified. This is also the case for the 3 best runs 4, 3 and 2 from the IBM AU were the performances reached impressive scores over 0.6 on the quality categories 4 & 5, 0.657 being the highest score obtained by IBM AU Run 4 on the quality category 4. It is interesting to notice that the scores of IBM AU Run 4 on the lowest quality categories 1 & 2 is competitive, even better, than the scores obtained on the highest quality categories 4 & 5

obtained by half of the submitted runs, which is highlighting the relevance of the approach developed by IBM AU (now acting as the new state of the art in plant identification).

3 Man vs. Machine experiment

3.1 Evaluation protocol

To make the benchmark reachable for human expertise we first drastically reduced the number of observations to be identified. From the 8163 observations in the test set of PlantCLEF we actually only kept 100 of them so that a human run typically took between 45 minutes and 2 hours, according to the person's level of expertise. The selection of these 100 queries was done by iteratively selecting 100 random observations from the whole dataset until the scores obtained by the LifeCLEF runs on that subset respect the ranking observed in the complete benchmark. This reduced test set was then shared with a large audience of potential volunteers composed of three target groups: skilled people (botanists, naturalists, teachers, more or less expert of the specifically targeted flora), amateurs (people interested by plants in parallel of their professional activity and having a knowledge of them at different expertise levels), novices (inexperienced users). From the hundreds of contacted people, 20 of them submitted a run (8 from the expert group, 7 from the amateur/student group, 5 from the novice group). Note that the novice group has been the less responding one mainly because of the higher difficulty and hardship for them to complete the task.
The human runs themselves were collected through a user interface presenting the 100 observations one by one (with one or several pictures of the different organs) and allowing the user to select up to three species for each observation thanks to a drop-down menu covering the 500 species of the PlantCLEF dataset. If the user provides no answer for a given query observation, the rank of the correct species is considered as infinite in the evaluation metric ($\frac{1}{r_{u,p=0}}$ in Equation 1). considered as infinite in the evaluation metric (formula 1). Note that to allow a fair comparison with the machines we also limited the machine runs to the same 100 observations and to the top-3 answers before computing the score S. It was also decided to display to the user the most popular common names of each of the 500 species in addition to the scientific name of the taxon, in order to facilitate participation of amateurs and novices. Date and location of botanical observations were also provided in each file name. As such information were also present in the xml files of the training and test observations, the machine runs could also have used it and this does not affect the fairness of the comparison.
A discussion has been conducted on whether we should provide pictures illustrating the species to be identified and/or or authorize the use of external resources such as field guides or the world wide web. A first positive argument was that the machines make use of the images in the training set so

Subject expertise	Avg nb of selected species	Avg nb of empty responses
Novices	1,27	70,6
Amateurs	1,52	38,57
Experts	1,28	13,13

Table 2 Statistics on the number of species selected by the 3 categories of subject – the average number of selected species is computed only on the non empty answers

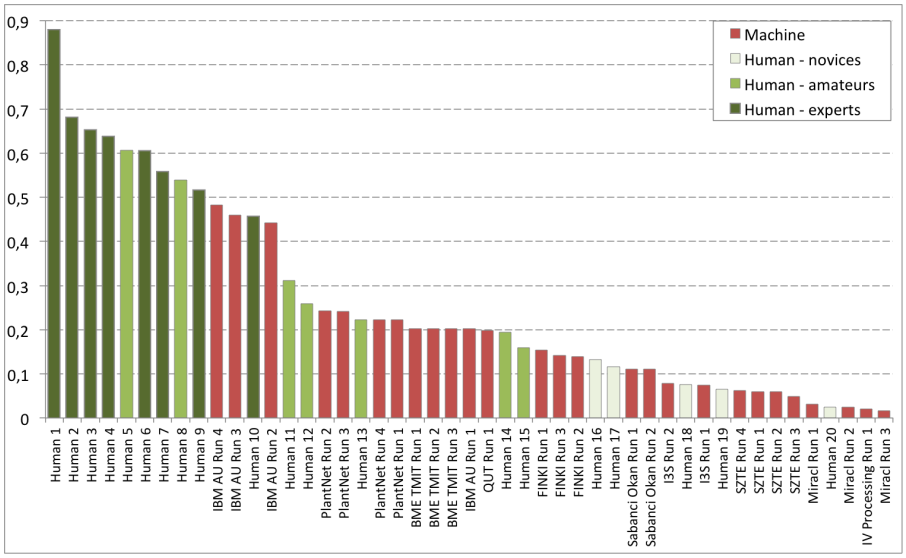

Fig. 5 Identification performances between humans (green) and machines (red), the y-axis is the metric defined in Equation (1).

that it would be fair to authorize humans to do so as well. But on the other side, this could favor the brute-force approach consisting in scanning all the species in the dataset one by one for each query which does not correspond to any real usage and which would have taken days for conscientious people. We therefore preferred restricting the evaluation to the knowledge-based identification of plants, without any additional information or tools during the test. As for the machines, people were thus supposed to have been trained beforehand. Concerning the use of external sources of information (such as data from wikipedia, eol[6] or ImageNET[7]), we remind that it was also forbidden for the machine runs of the LifeCLEF campaign so that the comparison appeared to be fair enough in that way.

[6] http://eol.org/
[7] http://www.image-net.org/

3.2 Results analysis

Table 3.1 first presents some statistics about the average number of species selected by the 3 categories of subjects. It shows that the average number of empty responses is clearly correlated with the degree of expertise. This means that the task is most of the time considered too hard by the subjects of the novice group so that they even don't try to provide an answer. Yet, when they decide to provide an answer to a given query, the average number of species they selected is roughly the same than the expert group. Only the amateur group tends to select more species probably because they only have a rough idea of what the plant is or is not.

Now, the identification results of the Man vs. Machine experiment are provided within Figure 5 with the machine runs in red (on the same subset of 100 queries) and the human runs in 3 different shades of green depending on the user's category (skilled, amateur or novice). First of all we can clearly note that no human was able to correctly identify all botanical observations even if a few of them are among the best known experts on French flora. This means that even for specialists themselves, the correct identification of botanical observation based on a few number of images is not always possible. The provided observations are actually most of the time incomplete for an accurate identification that would require more pictures of other parts of the plant. Also, botanical expertise is often organised by geographical zones or taxonomical groups. It is, in fact, extremely difficult to develop a strong expertise on large area and in the same time for a wide diversity of species. Our benchmark with species belonging to 91 plant families and distributed on the whole French territory is probably already too large to be fully succeeded by an expert. Furthermore, for the same species, plant morphology can vary a lot according to its growing stage, environmental condition or sexuality (for flowering plants, male and female flowers can be very different, in size, color or structure). Then, even if an expert knows the possible variations of a plant morphology according to his field experience or the literature, numerous information about this variability may not yet be recorded and shared through the scientific community, which can explain the difficulty to identify a species in all conditions.

Now, the main outcome of the evaluation is that the best human runs still clearly outperform all machine runs. All skilled people and several amateurs did correctly identify more observations than the best system of IBM. The worse performing systems have comparable performances with novices and inexperienced users which limit their practical interest. These conclusions can however be mitigated by a deeper analysis of the results with regard to the image quality. We therefore re-computed the per-quality projection of the human vs. machines runs as we did before for the official benchmark (see section 2.3). Figure 6 first displays the per-quality score of each of the human run. We can see that contrary to what we obtained on Figure 4 for the machine runs, human runs are much less affected by the quality of the test pictures. It is even hard to find any stable correlation between the identification performances and

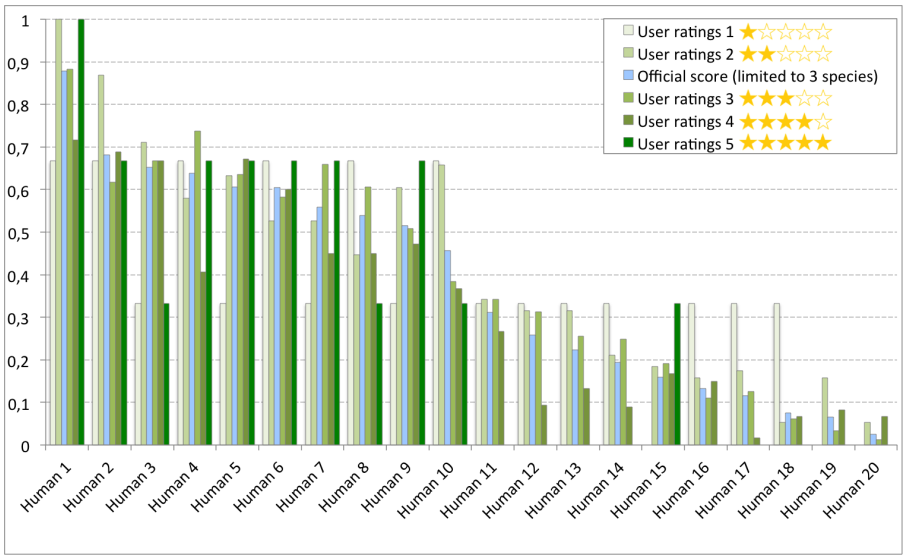

Fig. 6 Identification performances of the human runs detailed by quality categories

the quality of the observations. Very high-quality pictures with 5 stars ratings often lead to worse performance than 4, 3 or even 2 stars pictures for some users (independently from their expertise). This trend is confirmed by Figure 8 which displays the average identification performances across all human runs and all machine runs for the different categories of image quality. To further analyse the impact of image quality, Figure 7 displays the performances of both the machine and the human runs but restricted to the test observations with 3, 4 or 5 stars. It shows that the performances of the state-of-the-art systems are much closer to the human expertise if we do not consider the pictures with the lowest quality. IBM best run (based on Fisher vectors and support vector machines) notably provides competitive performances with expert botanists in that case. Most runs also provide much better performances than inexperienced users. Overall, that means that automated identification systems would be already functional for many users who are ready to spend a little time taking good pictures.

The results of our man vs. machine experiment finally have to be discussed with regard to the user perception of the evaluated identification approaches. Many of the test subjects, and especially the least expert in botany, did in particular complain about the hardship of the quiz itself. As they often did not know the name of the right species at the first glance, they sometimes tried to scan the 500 possible species and proceed by elimination before choosing one at random among the remaining candidates. This confirms the need of combining both automatic recognition algorithms and powerful interactive information retrieval mechanisms. As proved by a user study reported in [5], the

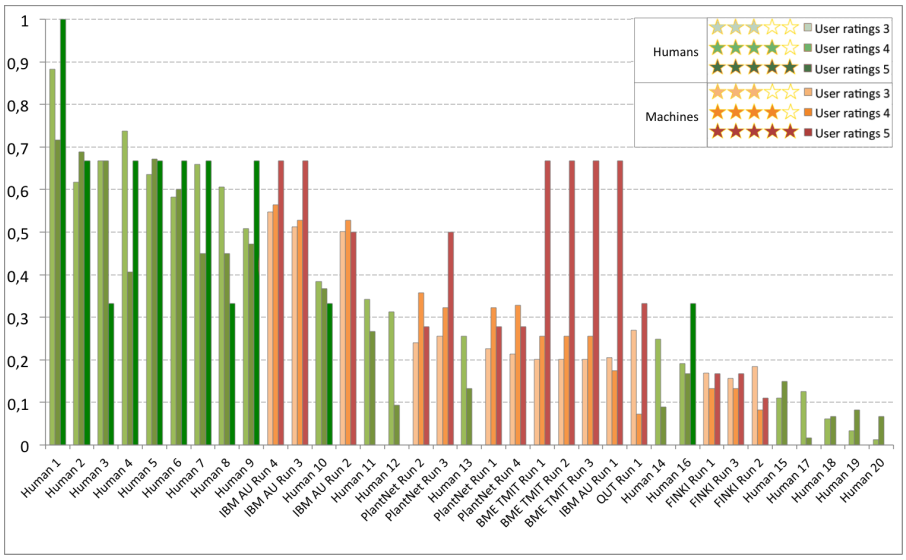

Fig. 7 Identification performances of the 15 best machine runs and 20 human runs restricted to high-quality pictures (with 3, 4 or 5 stars).

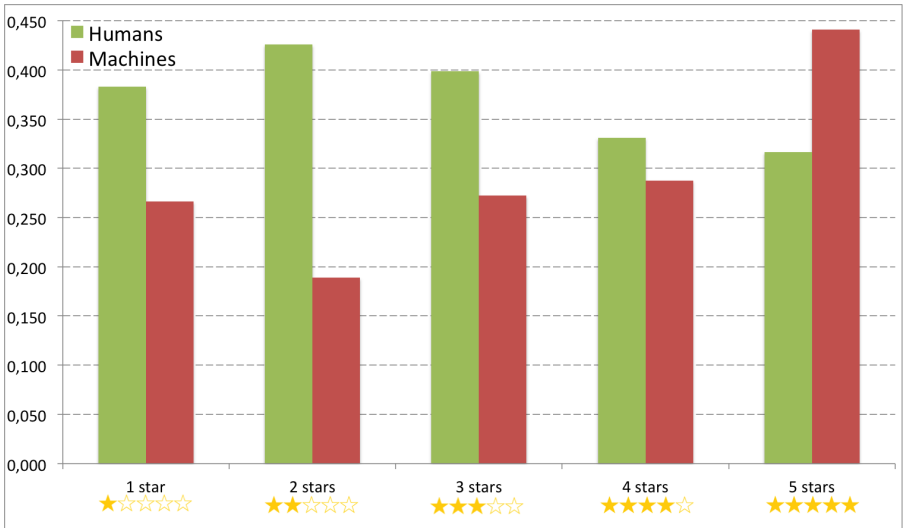

Fig. 8 Average identification performances across all human runs and all machine runs for the different categories of image quality.

raw performances of automatic identification systems can actually be strongly boosted once integrated in a real interactive application allowing the user to navigate the top results and visually check the proposed species. A perspective of the experiment reported in that paper could therefore be to complete

it with man+machine runs highlighting their mutual benefit.

4 Conclusion and perspectives

On May 11, 1997, Deep Blue, a chess-playing computer developed by IBM won the second six-game match against world champion Garry Kasparov, two to one, with three draws (with human intervention between games). At the image-based plant identification competition however, machines are still far from outperforming the best expert botanists. Our study shows that all skilled subjects with a minimal expertise in botany achieve better identification performances than the best systems evaluated within LifeCLEF 2014 (even if they were not expert of the specific flora addressed in the benchmark). We notably did show that the machine predictions are much more affected by the quality of the test images meaning that the underlying computer vision algorithms are still much more sensitive to the acquisition conditions than the human visual system. On the other side, the performances of the best systems are more than promising for the future. IBM best run (based on Fisher vectors and support vector machines) provides competitive performances with expert botanists on high-quality pictures. Several runs also provide much better performances than inexperienced users when restricting the evaluation to images with 3, 4 and 5 stars ratings. That means that semi-automated identification systems are already functional for users who are ready to spend a little time taking good pictures. Finally, it is important to remind that the use of external data was not authorized for training the systems evaluated in LifeCLEF. It is for instance worth noting that deep artificial neural networks could provide much better performances if they were pre-trained on large external corpora (even not related to botany). For the upcoming LifeCLEF evaluations, we will study the feasibility of providing more species as well as the feasibility of authorizing any other external training data to further improve the reliability of our challenge.

Acknowledgements Part of this work was funded by the Agropolis foundation through the project Pl@ntNet (`http://www.plantnet-project.org/`).

References

1. Gaston, Kevin J. and O'Neill, Mark A. Automated species identification: why not? Philosophical Transactions of the Royal Society B: Biological Sciences, volume 359, number 1444, pages 655-667, 2004.
2. Jinhai Cai, Ee, D., Binh Pham, Roe, P. and Jinglan Zhang, Sensor Network for the Monitoring of Ecosystem: Bird Species Recognition. 3rd International Conference on Intelligent Sensors, ISSNIP 2007.
3. Trifa, Vlad M., Kirschel, Alexander NG, Taylor, Charles E. and Vallejo, Edgar E. Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. The Journal of the Acoustical Society of America, volume 123, page 2424, 2008.

4. Spampinato and al. MAED '12: Proceedings of the 1st ACM International Workshop on Multimedia Analysis for Ecological Data. Nara, Japan. 2012.
5. Alexis Joly, Hervé Goëau, Pierre Bonnet, Vera Bakić, Julien Barbe, Souheil Selmi, Itheri Yahiaoui, Jennifer Carré, Elise Mouysset, Jean-François Molino, Nozha Boujemaa and Daniel Barthélémy. Interactive plant identification based on social image data. Ecological Informatics. 2013.
6. Nilsback, Maria-Elena and Zisserman, Andrew. Automated Flower Classification over a Large Number of Classes. Indian Conference on Computer Vision, Graphics and Image Processing. 2008.
7. Goëau, H., Joly, A. , Selmi, S., Bonnet, P., Mouysset, E., Joyeux, L., Molino, J-F., Birnbaum, P. Barthélémy, D. and Boujemaa, N. Visual-based plant species identification from crowdsourced data. ACM conference on Multimedia. 2011.
8. Cerutti, G., Tougne, L., Vacavant, A. and Coquin, D. A parametric active polygon for leaf segmentation and shape estimation. Advances in Visual Computing. pages 202-213, 2011.
9. Mouine, S., Yahiaoui, I. and Verroust-Blondet, A. Advanced shape context for plant species identification using leaf image retrieval. Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, 2012.
10. Kebapci, H., Yanikoglu, B. and Unal, G. Plant image retrieval using color, shape and texture features. The Computer Journal. 2010.
11. Hsu, T-H., Lee, C-H and Chen, L-H. An interactive flower image recognition system. Multimedia Tools and Applications, volume 53, number 1, pages 53-73, 2011.
12. Farnsworth, E. J., Chu, M., Kress, W. J., Neill, A. K., Best, J. H., Pickering, J., Stevenson, R. D., Courtney, G. W., VanDyk, J. K. and Ellison, A. M. Next-Generation Field Guides. volume 63, number 11, pages 891-899, 2013.
13. Neeraj Kumar, Peter N. Belhumeur, Arijit Biswas, David W. Jacobs, W. John Kress, Ida C. Lopez, João V. B. Soares. Leafsnap: A Computer Vision System for Automatic Plant Species Identification. European Conference on Computer Vision. 2012.
14. Goëau, H., Bonnet, P., Joly, A., Bakić, V., Barbe, J., Yahiaoui, I., Selmi, S., Carré, J., Barthélémy, D. and Boujemaa, N., PlantNet mobile app. Proceedings of the 21st ACM international conference on Multimedia, 2013.
15. Goëau, H., Bonnet, P., Joly, A., Affouard, A., Bakic, V., Barbe, J., Dufour, S., Selmi, S., Yahiaoui, I. and Vignau, C. PlantNet Mobile 2014: Android port and new features. Proceedings of International Conference on Multimedia Retrieval. 2014.
16. Goëau, H., Bonnet, P., Joly, A., Boujemaa, N., Barthélémy, D., Molino, J-F., Birnbaum, P., Mouysset, E. and Picard, M. The ImageCLEF 2011 plant images classification task. CLEF working notes, 2011.
17. Hervé Goëau, Pierre Bonnet, Alexis Joly, Itheri Yahiaoui, Daniel Barthélémy, Nozha Boujemaa and Jean-François Molino. The ImageCLEF 2012 Plant Identification Task. CLEF working notes, 2012.
18. Joly, A., Goëau, H., Bonnet, P., Bakic, V., Molino, J-F., Barthélémy, D. and Boujemaa, N. The Imageclef Plant Identification Task 2013. International workshop on Multimedia analysis for ecological data, Barcelone, Spain, 2013.
19. Goëau, H., Joly, A., Bonnet, P., Molino, J-F., Barthélémy, D. and Boujemaa, N. LifeCLEF Plant Identification Task 2014. CLEF working notes, 2014.
20. André Ricardo Backes, Dalcimar Casanova and Odemir Martinez Bruno. Plant Leaf Identification Based on Volumetric Fractal Dimension. International Journal of Pattern Recognition and Artificial Intelligence, volume 23, number 6, pages 1145-1160, 2009.
21. Guillaume Cerutti, Laure Tougne, Antoine Vacavant and Didier Coquin. A Parametric Active Polygon for Leaf Segmentation and Shape Estimation. International Symposium on Visual Computing. 2011.
22. Szúcs, G., Papp D. and Lovas, D. Viewpoints combined classification method in image-based plant identification task. Working notes of CLEF 2014 conference.
23. Karamti, H., Fakhfakh, S., Tmar, M. and Gargouri, F. MIRACL at LifeCLEF 2014: Multi-organ observation for Plant Identification. Working notes of CLEF 2014 conference.
24. Goëau, H., Joly, A., Yahiaoui, I., Bakić, V. and Verroust-Blondet A. PlantNet's participation at LifeCLEF 2014 Plant Identification Task. Working notes of CLEF 2014 conference.

25. Sunderhauf, N. McCool, C., Upcroft, B. and Perez, T. Fine-Grained Plant Classification Using Convolutional Neural Networks for Feature Extraction. Working notes of CLEF 2014 conference.
26. Chen, Q., Abedini, M., Garnavi, R., and Liang, X. IBM Research Australia at LifeCLEF2014: Plant Identification Task. Working notes of CLEF 2014 conference.
27. Yanikoglu, B., S. Tolga, Y., Tirkaz, C. , and FuenCaglartes, E. Sabanci-Okan System at LifeCLEF 2014 Plant Identification Competition. Working notes of CLEF 2014 conference.
28. Dimitrovski, I., Madjarov, G., Lameski, P. and Kocev, D. Maestra at LifeCLEF 2014 Plant Task: Plant Identification using Visual Data. Working notes of CLEF 2014 conference.
29. Paczolay, D., Bánhalmi, A., Nyúl, L., Bilicki, V. and Sárosi, Á. Wlab of University of Szeged at LifeCLEF 2014 Plant Identification Task. Working notes of CLEF 2014 conference.
30. Fakhfakh, S., Akrout, B., Tmar, M. and Mahdi, W. A visual search of multimedia documents in LifeCLEF 2014. Working notes of CLEF 2014 conference.
31. Issolah, M., Lingrand, D., and Precioso, F. Plant species recognition using Bag-Of-Word with SVM classifier in the context of the LifeCLEF challenge. Working notes of CLEF 2014 conference.
32. Joly, A., Müller, H., Goëau, H., Glotin, H., Spampinato, C., Rauber, A., Bonnet, P., Vellinga, W-P. and Fisher, B. LifeCLEF 2014: multimedia life species identification challenges. Proceedings of CLEF 2014.
33. Anelia Angelova, Shenghuo Zhu, Yuanqing Lin, Josephine Wong and Chelsea Shpecht. Development and Deployment of a Large-scale Flower Recognition Mobile App. NEC Labs America Technical Report, 2012.
34. Voorhees, E. M. The TREC-8 Question Answering Track Report. TREC (Vol. 99, pp. 77-82), 1999.